Annahmeverletzungen und Schätzabweichungen des Rasch-Modells

Masterarbeit eingereicht bei

Prof. Dr. Wolfgang K. Härdle

Dr. Sigbert Klinke

Ladislaus von Bortkiewicz Lehrstuhl für Statistik

C.A.S.E. – Center for Applied Statistics and Economics

Humboldt-Universität zu Berlin

von

Christoph Jaehrling

(Matrikel-Nr. 542214)

zur Erlangung des akademischen Grades eines

Master of Science in Statistik

Berlin, 9. September 2013

Zusammenfassung

Im Mittelpunkt der Arbeit steht die Quantifizierung von Schätzabweichungen der Personen- und Aufgabenparameter im Rasch-Modell, insbesondere unter Berücksichtigung von Modellverletzungen. Mithilfe einer Monte-Carlo-Simulation werden „wahre" Parameter generiert und eine Antwortmatrix erzeugt. Daraus berechnet das Rasch-Modell Schätzwerte für die Fähigkeiten von Personen und Schwierigkeiten von Aufgaben. Von den Schätzungen und den wahren Werten werden die mittleren betragsmäßigen Abweichungen und die Bias bestimmt. Die Einflüsse der Modellverletzungen auf die Schätzabweichungen werden mithilfe von robusten Regressionen (mit Bootstrapping) evaluiert. Die Analyse zeigt, dass die Parameter bei Gültigkeit des Rasch-Modells erwartungstreu geschätzt werden. Bei Annahmeverletzungen steigen die Schätzabweichungen zum Teil erheblich. Als schwerwiegendste Verletzung stellt sich die lokale stochastische Abhängigkeit heraus.

Schlagwörter: Rasch-Modell, Schätzabweichungen, bedingte Maximum-Likelihood-Schätzung, Monte-Carlo-Simulation, robuste Regression

Abstract

Focus of the thesis is the quantification of estimation errors of the person and item parameters in the Rasch model, under special consideration of model violations. With a Monte Carlo simulation "true" parameters are generated and a response matrix is created. From that matrix the Rasch model estimates the abilities of persons and the difficulties of items. Based on the estimates and the true values, mean absolute deviations and biases are computed. The effects of model violations on estimation errors are evaluated using robust regressions (with bootstrapping). The analysis shows that the parameter estimates are unbiased under validity of the Rasch model. When the assumptions are violated the estimation errors increase considerably in most cases. Local stochastic dependence is considered as the most serious violation.

Keywords: Rasch model, estimation errors, conditional maximum likelihood estimation, Monte Carlo method, robust regression

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Das Rasch-Modell ist ein Spezialfall (und Vorläufer) der Item-Response-Theorie. Es trägt seinen Namen nach dem dänischen Statistiker Georg Rasch (1960). Dieser präsentierte vor 50 Jahren ein Messmodell, welches (in einem Test mit dichotomen Aufgaben) einen probabilistischer Zusammenhang zwischen dem Antwortverhalten eines Befragten und einer interessierenden latenten Variablen herstellt. Ziel ist es, damit Fähigkeiten von Personen und Schwierigkeiten von Aufgaben zu schätzen. Das Rasch-Modell wird überwiegend für Leistungstests eingesetzt, etwa dem HAM-Nat, einem Auswahlverfahren für die Studiengänge Human- und Zahnmedizin an der Universität Hamburg.

Dem Modell liegen im Wesentlichen drei Annahmen zugrunde: Spezifische Objektivität, Eindimensionalität und lokale stochastische Unabhängigkeit. Für die Gültigkeit des Rasch-Modells wird zwar gefordert, dass die Annahmen unverletzt sein sollen. In der Praxis (vor allem bei der Entwicklung von Aufgaben) ist es jedoch schwierig sie nicht zu verletzen, ehe man sie getestet hat.[1] Darüber hinaus ist es wissenswert zu quantifizieren, welche Schätzabweichungen mit den Annahmeverletzungen einhergehen.

Auch ohne Modellverletzungen unterscheiden sich die ermittelten Fähigkeiten derselben Personen in wiederholten Messungen häufig (unter gleichen Bedingungen). Dabei gibt es zufällige und systematische Abweichungen. Letztere sind Fehler, die die Messwerte (bei sehr vielen Messungen) entweder über- oder unterschätzen. Um Messabweichungen für Personenfähigkeiten (etwa Intelligenzen) quantifizieren zu können, müssten die wahren Fähigkeiten bekannt sein. Wären sie aber bekannt, müssten sie nicht geschätzt werden. Deshalb ist es sehr schwierig, Schätzabweichungen des Rasch-Modells empirisch zu evaluieren. Es ist aber möglich *echte* Werte mit Hilfe von Monte-Carlo-Simulationen zu generieren, das Rasch-Modell diese schätzen zu lassen und davon die Abweichungen zu ermitteln. Das ist das Konzept dieser Arbeit.

In Kapitel 2 werden zunächst die Grundlagen des Rasch-Modells – der Modellrahmen, die zentralen Annahmen und die Schätzung der Parameter – skizziert. In Kapitel 3 wird das Simulationsdesign erläutert, vor allem die Datengenerierung und die Berücksichtigung der Annahmeverletzungen. In Kapitel 4 werden die zur Quantifizierung der Schätzabweichungen notwendigen Abweichungsmaße sowie die zur Beurteilung der Einflussfaktoren verwendete Regressionsanalyse beschrieben. Schließlich werden die Ergebnisse in Kapitel 5 vorgestellt und bewertet.

[1] Es konnte bereits gezeigt werden, mit welchen Teststatistiken sich Verletzungen dieser Annahmen entdecken lassen (vgl. etwa Suárez-Falcón und Glas 2003 oder Mair und Ledl 2006).

2 Grundlagen des Rasch-Modells

Das Rasch-Modell ist ein Instrument für Leistungs- oder Persönlichkeitstests, in dem vom Antwortverhalten der Befragten auf ihre Fähigkeiten bzw. Neigungen hinsichtlich einer latenten Variablen geschlossen wird. Gleichzeitig können damit die Schwierigkeiten (und die Brauchbarkeit) der Aufgaben ermittelt werden, womit sie adaptives Testen ermöglichen.

Die Aussagen in diesem Kapitel beziehen sich überwiegend auf die Arbeiten von Strobl (2012) und Rost (2004), sowie auch auf Höhne und Hölzlwimmer (2009) für den Abschnitt der Parameterschätzung (2.3).

2.1 Modellrahmen

Man stelle sich vor, dass $i = 1, \ldots, n$ Personen an einem Test mit $j = 1, \ldots, m$ Aufgaben (Items) teilnehmen. Die Antworten der Befragten ließen sich in einer Datenmatrix zusammenfassen, welche (in allgemeiner Form) in Tabelle 2.1 dargestellt ist.

	Aufgabe						
Person	1	2	$\cdots$	j	$\cdots$	$m-1$	m
1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,j}$	$\cdots$	$x_{1,m-1}$	$x_{1,m}$
2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,j}$	$\cdots$	$x_{2,m-1}$	$x_{2,m}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
i	$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,j}$	$\cdots$	$x_{i,m-1}$	$x_{i,m}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
$n-1$	$x_{n-1,1}$	$x_{n-1,2}$	$\cdots$	$x_{n-1,j}$	$\cdots$	$x_{n-1,m-1}$	$x_{n-1,m}$
n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,j}$	$\cdots$	$x_{n,m-1}$	$x_{n,m}$

Tabelle 2.1: Datenmatrix eines Tests mit n Personen und m Aufgaben

Dabei entspricht $x_{i,j}$ der Antwort der i-ten Person auf die j-te Frage.

Eine Antwort hat nur zwei mögliche Ausprägungen. Sie ist entweder *richtig* (1) oder *falsch* (0).[1] Ehe die Befragten eine Aufgabe bearbeiten, ist die spätere tatsächliche Antwort noch unbekannt und wird in diesem Fall als Zufallsvariable $X_{i,j}$ beschrieben.

Die Wahrscheinlichkeit für einen Befragten i eine Frage j richtig zu lösen, ist auf den Wertebereich $[0, 1]$ beschränkt und wird durch eine logistische Funktion, der Rasch-Modellgleichung,

[1] Hierbei wird vom Kontext eines Leistungstest, z.B. in Form eines Intelligenztests, ausgegangen. In anderen Anwendungen könnten die beiden Ausprägungen etwa auch *Zustimmung* und *Ablehnung* lauten.

beschrieben:

$$\mathcal{P}(X_{i,j} = 1|\theta_i, \delta_j) = \frac{\exp\{\theta_i - \delta_j\}}{1 + \exp\{\theta_i - \delta_j\}} \tag{2.1}$$

Die Lösungswahrscheinlichkeit hängt dabei sowohl von der Personenfähigkeit θ_i als auch der Aufgabenschwierigkeit δ_j ab. Der Wertebereich beider Parameter liegt theoretisch zwischen $-\infty$ und $+\infty$, in der Praxis aber meist zwischen -3 und 3. Je größer θ_i, desto höher ist die Fähigkeit der Person i. Je größer δ_j, desto schwieriger ist die Aufgabe j. Beide Parameter müssen die gleiche Einheit besitzen, um sinnvoll verrechnet zu werden. (vgl. Höhne und Hölzlwimmer 2009: 2)

Die Modellgleichung lässt sich auch als aufgabencharakteristische Kurve (*item characteristic curve*, ICC) darstellen. Abbildung 2.1 zeigt ein Beispiel einer solchen Kurve für eine Aufgabe mit Schwierigkeit $\delta_j = 0$.

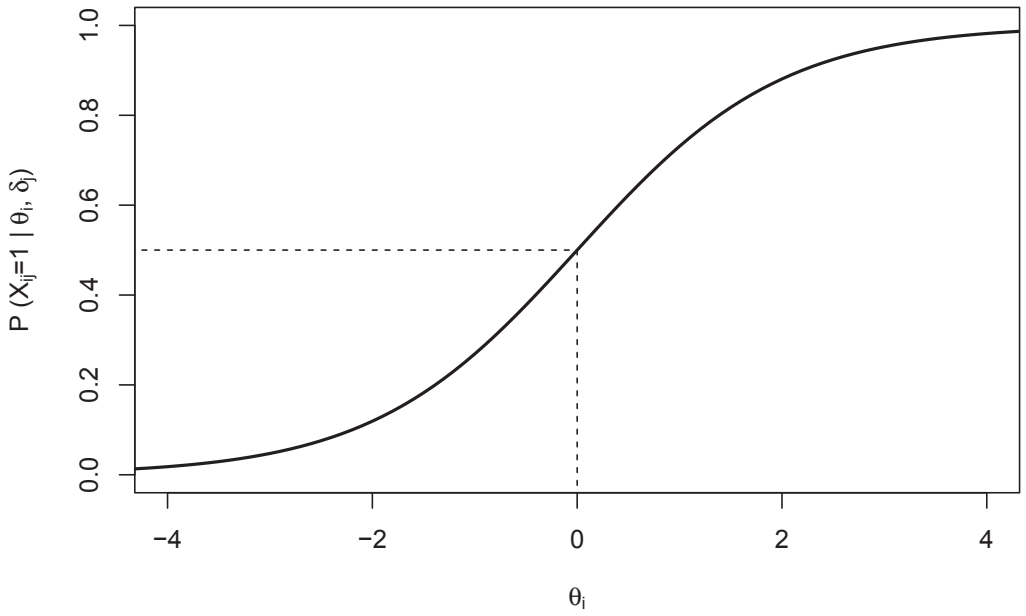

Abbildung 2.1: ICC für eine Aufgabe mit Schwierigkeit $\delta_j = 0$

Die Funktion steigt, wenn der Exponent $\theta_i - \delta_j$ wächst. Entsprechen beide Parameter einander (etwa wie im Fall von Abbildung 2.1 mit $\theta_i = \delta_j = 0$) liegt die Lösungswahrscheinlichkeit bei 0,5. Bei „Überlegenheit“ der Person über die Aufgabe ($\theta_i > \delta_j$) ist die Wahrscheinlichkeit für eine richtige Antwort größer als 0,5.

2.2 Zentrale Annahmen

2.2.1 Spezifische Objektivität

Georg Rasch bezeichnet die Eigenschaft, die grafisch durch die parallelen ICCs veranschaulicht wird – die lediglich nach links oder rechts verschoben sind – als spezifische Objektivität (siehe Abbildung 2.2). Sie impliziert, dass Vergleiche der Fähigkeiten zweier Personen im Rasch-Modell aufgabenunabhängig sind, weil sich jede Aufgabe dafür gleichermaßen eignet. Die Annahme gilt auch für den Vergleich von Aufgaben. Leichtere Fragen (in Abbildung 2.2 eher links) haben unabhängig von den Personenfähigkeiten immer höhere Lösungswahrscheinlichkeiten als schwierige.

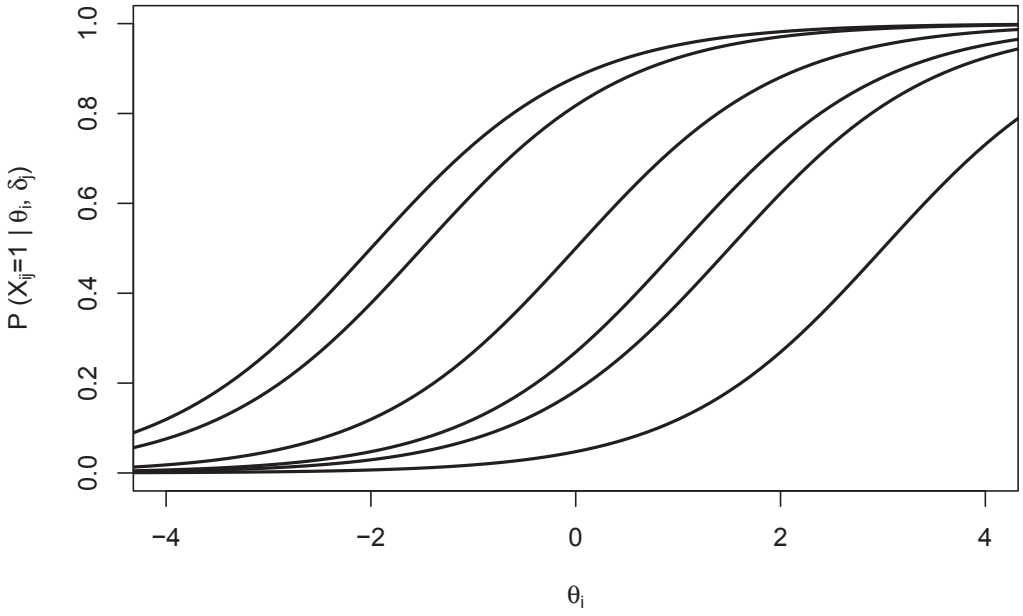

Abbildung 2.2: ICCs für Aufgaben mit unterschiedlichen Schwierigkeiten

Würden sich die ICCs hingegen schneiden, so wäre eine richtige Antwort für Befragte mit hohen Fähigkeiten zwar nach wie vor in jeder Frage wahrscheinlicher als für Befragte mit geringen Fähigkeiten (solange der Anstieg positiv bliebe). Letztere hätten allerdings in einer Situation wie in Abbildung 2.3 die höchsten Lösungswahrscheinlichkeiten für die schwierigere Frage 2 ($\delta_2 = 2$), weil diese Aufgabe weit weniger zwischen den Befragten diskriminiert als die leichtere Frage 1 ($\delta_1 = 0$). Aufgabenvergleiche wären dann nicht personenunabhängig und Personenvergleiche nicht aufgabenunabhängig. Die spezifische Objektivität wäre also verletzt.

Im Birnbaum-Modell Birnbaum (1968) – einer Erweiterung des Rasch-Modells – gibt es

einen dritten Parameter, der unterschiedliche Steigungen zulässt. Durch diesen *Diskriminationsparameter* (auch als Trennschärfe bezeichnet) wirken sich Unterschiede der Personenfähigkeiten (bei gleicher Aufgabenschwierigkeit) feiner auf die Lösungswahrscheinlichkeit aus.

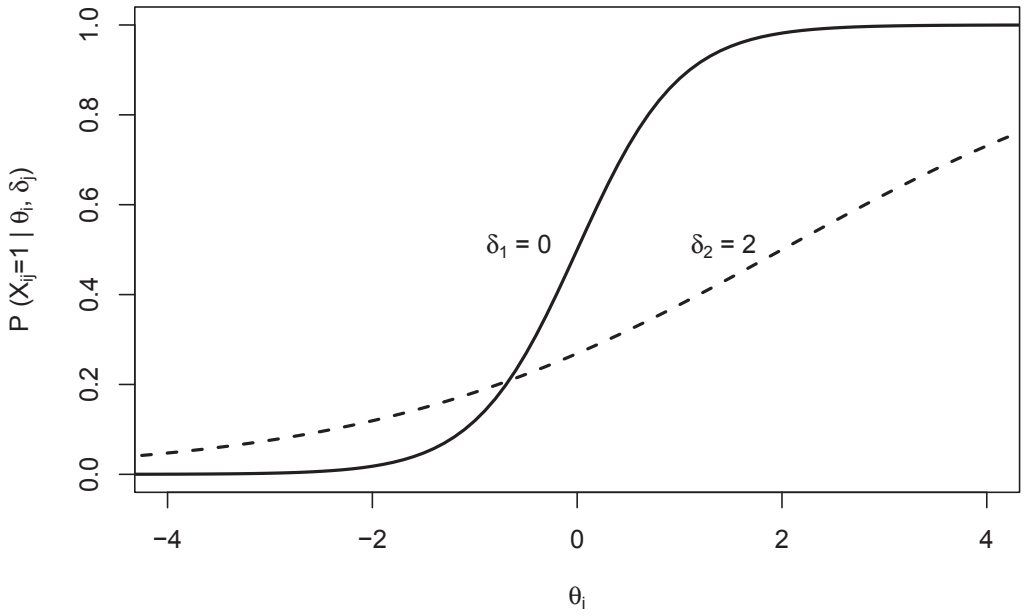

Abbildung 2.3: Beispiel für eine Verletzung der spezifischen Objektivität

2.2.2 Eindimensionalität

Alle Personen- und Aufgabenparameter im Modell beziehen sich nur auf ein gemeinsames latentes Konstrukt. In der Modellgleichung ist dies dadurch gekennzeichnet, dass die Parameter θ_i und δ_j direkt voneinander subtrahiert werden. Grafisch wird deutlich, dass beide auf der x-Achse abgetragen werden (siehe etwa Abbildung 2.3).

Die Annahme der Eindimensionalität wäre verletzt, wenn neben einer latenten Variablen noch eine weitere gemessen würde. Exemplarisch könnte man sich hier einen Mathematik-Test vorstellen, in dem Textaufgaben vorkommen, die neben mathematischen Kenntnissen auch fortgeschrittene Sprachkompetenzen voraussetzen.

Mit den mehrdimensionalen Rasch-Modellen existieren Erweiterungen des Basismodells, die sich zur Messung unterschiedlicher Zielvariablen eignen (etwa Glas 1992 oder Carstensen und Rost 2011).

2.2.3 Lokale stochastische Unabhängigkeit

Die Lösungswahrscheinlichkeiten für die einzelnen Aufgaben sollen stochastisch unabhängig voneinander sein. Die Wahrscheinlichkeit für eine richtige Antwort soll nur von der betreffenden Aufgabe abhängen und nicht von der Lösung einer anderen Aufgabe. Das gilt jedoch nur für die Betrachtung einzelner Personen und heißt deshalb lokal. Befragte mit höheren Fähigkeiten haben zwar für alle Fragen höhere Lösungswahrscheinlichkeiten als Befragte mit geringeren Fähigkeiten, aber eine Aufgabe darf nicht auf eine andere aufbauen.

Diese Annahme erleichtert vor allem die Berechnung gemeinsamer Wahrscheinlichkeiten von Ereignissen, da sie sich bei stochastischer Unabhängigkeit als Produkt der einzelnen Eintrittswahrscheinlichkeiten ergeben. Eine Verletzung dieser Annahme führt zu verzerrten Schätzungen der Fähigkeiten und Aufgaben (vgl. Wang und Wilson 2005: 126).

Erweiterungen des Rasch-Modells, die zusammenhängende Aufgaben berücksichtigen, sind etwa Testlet-Modelle (vgl. Wainer und Kiely 1987 für IRT-Modelle im Allgemeinen sowie Wang und Wilson 2005 für das Rasch-Modell als Spezialfall). Diese enthalten einen zusätzlichen Faktor, der die Abhängigkeit von verbundenen Items modelliert.

2.3 Parameterschätzung

In der Rasch-Modellgleichung (Formel 2.1) kommen mit der Personenfähigkeit θ_i und der Aufgabenschwierigkeit δ_j zwei Parameter vor, die zunächst unbekannt sind und aus einer Datenmatrix wie der in Tabelle 2.1 geschätzt werden müssen. Dazu gibt es unterschiedliche Ansätze, die jedoch alle auf den Maximum-Likelihood-Ansatz aufbauen. Im Folgenden wird nur die bedingte Maximum-Likelihood-Schätzung erläutert, weil sie – im Gegensatz zur gemeinsamen Maximum-Likelihood-Schätzung – konsistente Schätzer liefert und am weitesten verbreitet ist. Sie ist auch im R-Paket `eRm` implementiert, welches für die späteren Berechnungen genutzt wird.

Unter der Annahme, dass das Modell gilt, beschreibt die Likelihoodfunktion die Wahrscheinlichkeit der Daten unter Berücksichtigung der Modellparameter. (vgl. Rost 2004: 112)

Die Likelihoodfunktion $\mathcal{L}$ für eine Person i und eine Aufgabe j ist:

$$\mathcal{L}_{x_{i,j}}(\theta_i, \delta_j) = \mathcal{P}(X_{i,j} = x_{i,j} | \theta_i, \delta_j) = \frac{\exp\{x_{i,j}\left(\theta_i - \delta_j\right)\}}{1 + \exp\{\theta_i - \delta_j\}} \tag{2.2}$$

Daraus sollen zunächst die Aufgabenparameter δ_j bestimmt werden. Für eine Person i über

alle Aufgaben $j = 1, \ldots, m$ ist die Likelihoodfunktion:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x_i}}(\theta_i, \boldsymbol{\delta}) = \mathcal{P}(\boldsymbol{X_i} = \boldsymbol{x_i}|\theta_i, \boldsymbol{\delta}) &= \prod_{j=1}^{m} \mathcal{P}(X_{i,j} = x_{i,j}|\theta_i, \delta_j) \\ &= \prod_{j=1}^{m} \frac{\exp\{x_{i,j}\left(\theta_i - \delta_j\right)\}}{1 + \exp\{\theta_i - \delta_j\}} \\ &= \frac{\exp\{\sum_{j=1}^{m} x_{i,j}\left(\theta_i - \delta_j\right)\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]} \\ &= \frac{\exp\{\sum_{j=1}^{m} x_{i,j} \cdot \theta_i - \sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]}\end{aligned} \tag{2.3}$$

Hierbei sind $\boldsymbol{X_i}$, $\boldsymbol{x_i}$ und $\boldsymbol{\delta}$ Vektoren, welche die verfügbaren Informationen von Person i enthalten. Der Ausdruck $\sum_{j=1}^{m} x_{i,j}$ lässt sich auch als Summenscore (oder Personen-Randsumme) über alle Antworten eines Befragten interpretieren und vereinfacht als r_i schreiben, sodass:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x_i}}(\theta_i, \boldsymbol{\delta}) &= \frac{\exp\{r_i \cdot \theta_i - \sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]} \\ &= \frac{\exp\{r_i \cdot \theta_i\} \cdot \exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]}\end{aligned} \tag{2.4}$$

Mithilfe der Summenscores r_i lässt sich die Likelihoodfunktion faktorisieren:

$$\mathcal{L}_{\boldsymbol{x_i}}(\theta_i, \boldsymbol{\delta}) = h(\boldsymbol{x_i}|r_i, \theta_i, \boldsymbol{\delta}) \cdot g(r_i|\theta_i, \boldsymbol{\delta}) \tag{2.5}$$

Davon ist $h(\boldsymbol{x_i}|r_i, \theta_i, \boldsymbol{\delta})$ die bedingte Likelihoodfunktion und $g(r_i|\theta_i, \boldsymbol{\delta})$ die Summe der Wahrscheinlichkeiten über alle Antwortvektoren $\boldsymbol{x_i}$ für einen bestimmten Summenscore r_i:

$$g(r_i|\theta_i, \boldsymbol{\delta}) = \sum_{\boldsymbol{x}|r_i} \frac{\exp\{r_i \cdot \theta_i\} \cdot \exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]} \tag{2.6}$$

Dann kann die Gleichung 2.5 nach $h(\boldsymbol{x_i}|r_i, \theta_i, \boldsymbol{\delta})$ umgeformt werden:

$$\begin{aligned}h(\boldsymbol{x_i}|r_i, \theta_i, \boldsymbol{\delta}) &= \frac{\mathcal{L}_{\boldsymbol{x_i}}(\theta_i, \boldsymbol{\delta})}{g(r_i|\theta_i, \boldsymbol{\delta})} \\ &= \frac{\frac{\exp\{r_i \cdot \theta_i\} \cdot \exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]}}{\sum_{\boldsymbol{x}|r_i} \frac{\exp\{r_i \cdot \theta_i\} \cdot \exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\prod_{j=1}^{m}\left[1 + \exp\{\theta_i - \delta_j\}\right]}} \\ &= \frac{\exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\sum_{\boldsymbol{x}|r_i} \exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}} \\ &= \mathcal{L}_{\boldsymbol{x_i}}(r_i, \boldsymbol{\delta})\end{aligned} \tag{2.7}$$

$\mathcal{L}_{\boldsymbol{x_i}}(r_i, \boldsymbol{\delta})$ ist die bedingte Likelihoodfunktion für Person i über alle Aufgaben – unabhängig

von der Personenfähigkeit θ_i. Die bedingte Likelihoodfunktion $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{r}, \boldsymbol{\delta})$ ist das Produkt über alle Befragten:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{r}, \boldsymbol{\delta}) &= \prod_{i=1}^{n} \mathcal{L}_{\boldsymbol{x}_i}(r_i, \boldsymbol{\delta}) \\ &= \prod_{i=1}^{n} \frac{\exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}{\sum_{\boldsymbol{x}|r_i} \exp\{-\sum_{j=1}^{m} x_{i,j} \cdot \delta_j\}}\end{aligned} \tag{2.8}$$

Um die Aufgabenparameter zu schätzen, wird die Likelihoodfunktion $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{r}, \boldsymbol{\delta})$ maximiert, indem sie logarithmiert, nach δ_j abgeleitet und das Ergebnis gleich 0 gesetzt wird. Da sich Gleichung 2.8 nicht nach δ_j auflösen lässt, verwendet man iterative Verfahren zur Schätzung. Dabei wird in der Regel festgelegt, dass sich die Aufgabenparameter zu 0 summieren.[2]

Nachdem man so die Aufgabenparameter erhalten hat, setzt man sie in die unbedingte Likelihoodfunktion $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{\theta}, \boldsymbol{\delta})$ ein und schätzt die Personenparameter:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{\theta}, \boldsymbol{\delta}) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\exp\{x_{i,j}\left(\theta_i - \delta_j\right)\}}{1 + \exp\{\theta_i - \hat{\delta}_j\}} \\ &= \frac{\exp\{\sum_{i=1}^{n} r_i \cdot \theta_i\} \cdot \exp\{-\sum_{j=1}^{m} s_j \cdot \hat{\delta}_j\}}{\prod_{i=1}^{n} \prod_{j=1}^{m} \left[1 + \exp\{\theta_i - \hat{\delta}_j\}\right]}\end{aligned} \tag{2.9}$$

Dabei ist $s_j = \sum_{i=1}^{n} x_{i,j}$. Auch diese Likelihoodfunktion $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{\theta}, \boldsymbol{\delta})$ wird maximiert, also logarithmiert, (nach θ_i) abgeleitet und gleich 0 gesetzt. Der Ausdruck lässt sich nach r_i umformen:

$$r_i = \sum_{j=1}^{m} \frac{\exp\{\theta_i - \hat{\delta}_j\}}{1 + \exp\{\theta_i - \hat{\delta}_j\}} \tag{2.10}$$

Hieraus lassen sich die Personenfähigkeiten θ_i iterativ schätzen. Dabei werden standardisierte Werte ausgegeben (Mittelwert $= 0$, Standardabweichung $= 1$).

Die bedingte Maximum-Likelihood-Schätzung hat den Nachteil, dass sich die Fähigkeitsparameter nicht für Personen bestimmen lassen, die entweder alle oder gar keine Aufgabe richtig gelöst haben. Das Gleiche gilt für die Schwierigkeitsparameter von Aufgaben, die entweder von allen Teilnehmern richtig oder von allen falsch beantwortet wurden. Solche Werte werden dann extrapoliert.

[2] Die Festlegung auf die Summe gleich 0 ist nicht erforderlich. Entscheidend ist, dass ein Identifikationsproblem vorliegt, weshalb nicht alle Parameter frei geschätzt werden können und ein Wert fixiert werden muss.

3 Simulationsdesign

Das Anliegen dieser Arbeit ist es Schätzabweichungen des Rasch-Modells für die Personenfähigkeiten und die Aufgabenschwierigkeiten zu quantifizieren. Da deren wahre Werte unbekannt sind, lässt sich diese Aufgabe nicht mithilfe von empirischen Daten bewältigen.

Das Design dieser Simulation orientiert sich an den Studien von Suárez-Falcón und Glas (2003) sowie Mair und Ledl (2006)[1]. Sie manipulieren die Rasch-Modellgleichung mit Annahmeverletzungen und evaluieren damit globale Modellanpassungstests hinsichtlich ihrer Eignung die Verletzungen aufzudecken.

Anders als in den Referenzstudien werden in dieser Arbeit keine Anpassungstests evaluiert, sondern die Schätzabweichungen für die Modellparameter θ_i und δ_j. Der Lösungsansatz sieht deshalb vor, wahre Fähigkeiten θ_i für künstlich erzeugte Befragte vorzugeben. Diese versuchen eine bestimmte Anzahl an hypothetischen Aufgaben zu lösen, deren wahre Schwierigkeiten δ_j ebenfalls festgelegt sind. Unter Verwendung der Antwortmuster schätzt das Rasch-Modell dann die Personenfähigkeiten und die Aufgabenschwierigkeiten, die mit den generierten wahren Werten verglichen werden.

Eine Monte-Carlo-Simulation bietet die Möglichkeit diesen Prozess mithilfe von vielfach durchgeführten Zufallsexperimenten zu simulieren. Für diese Arbeit werden 10.000 Durchgänge mit der Statistiksoftware R realisiert.

3.1 Generierung der wahren Werte

Der erste Schritt der Simulation ist die Bestimmung der Anzahl an Befragten n und Aufgaben m. Da frühere Simulationsstudien gezeigt haben, dass auch sie einen Einfluss auf die Schätzgüte des Rasch-Modells besitzen (etwa Suárez-Falcón und Glas sowie Mair und Ledl), sollen sie nicht in jedem Simulationsdurchlauf auf die gleichen Werte fixiert sein. Stattdessen werden sie für jeden Durchgang (also für insgesamt 10.000 hypothetische Tests) zufällig aus einer diskreten Gleichverteilung $\mathcal{U}_D$ erzeugt[2]:

$$n \sim \mathcal{U}_D[100, 1.000] \tag{3.1}$$

$$m \sim 2 \cdot \mathcal{U}_D[5, 50] \tag{3.2}$$

Das bedeutet, dass in jedem Test zwischen 100 und 1.000 Befragte 10 bis 100 Fragen beantworten, wobei n und m unabhängig voneinander und von anderen Simulationen sind.

Danach werden für alle n Befragten tatsächliche Fähigkeiten θ_i und für alle m Aufgaben

[1] Wobei sich Mair und Ledl ebenfalls auf das Studiendesign von Suarez-Falcon und Glas beziehen.

[2] Wegen der späteren Implementierung von Annahmeverletzungen muss m immer eine gerade Zahl sein.

wahre Schwierigkeiten δ_j (hier zunächst als $\dot{\delta}_j$ bezeichnet) zufällig aus einer Standardnormalverteilung generiert:

$$\theta_i \sim \mathcal{N}(0,1) \qquad \forall i = 1,\ldots,n \tag{3.3}$$
$$\dot{\delta}_j \sim \mathcal{N}(0,1) \qquad \forall j = 1,\ldots,m \tag{3.4}$$

Da in der späteren Schätzung die Summe der Aufgabenparameter auf 0 festgelegt wird, sich die generierten echten Aufgabenparameter aber nicht notwendigerweise zu 0 summieren, werden alle $\dot{\delta}_j$ eines Durchlaufs zentriert:

$$\delta_j = \dot{\delta}_j - \bar{\dot{\delta}} \qquad \forall j = 1,\ldots,m \tag{3.5}$$

Mit den definierten Parametern ist es bereits möglich eine Datenmatrix zu erzeugen, mit der das Rasch-Modell geschätzt werden kann. Im Zentrum der Arbeit steht aber die Quantifizierung von Schätzabweichungen, die aus Modellverletzungen resultieren.

3.2 Implementierung von Annahmeverletzungen

3.2.1 Verletzung der spezifischen Objektivität

Eine Verletzung der spezifischen Objektivität bedeutet formal, dass sich die aufgabenspezifischen Kurven schneiden. Dies ist in der Item-Response-Theorie generell erlaubt und entspricht dem 2PL-Modell nach Birnbaum (1968). Dafür wird ein zusätzlicher Term γ_j – welcher auch als Diskriminationsparameter bezeichnet wird – in die Modellgleichung (2.1) eingeführt:

$$\mathcal{P}(X_{i,j} = 1|\theta_i, \delta_j, \gamma_j) = \frac{\exp\{\gamma_j(\theta_i - \delta_j)\}}{1 + \exp\{\gamma_j(\theta_i - \delta_j)\}} \tag{3.6}$$

Der Diskriminationsparameter γ_j ist – wie der Schwierigkeitsparameter δ_j – an eine bestimmte Aufgabe j geknüpft. Deshalb wird er in jedem Durchlauf m-mal generiert:

$$\gamma_j \sim \ln\mathcal{N}(0, \sigma_\gamma^2) \qquad \forall j = 1,\ldots,m \tag{3.7}$$

Mit Bezug auf Suárez-Falcón und Glas wird γ_j aus einer logarithmischen Normalverteilung gezogen.[3] Die Standardabweichung σ_γ bestimmt dabei das Ausmaß der Verletzung der spezifischen Objektivität und wird im Folgenden mit ψ_{obj} bezeichnet. ψ_{obj} wird zufällig aus einer stetigen Gleichverteilung $\mathcal{U}_S$ mit den Grenzen dieser Extremwerte bestimmt (für jeden der 10.000 Durchläufe einmal):

$$\psi_{obj} = \sigma_\gamma \sim \mathcal{U}_S[0,0{,}5] \tag{3.8}$$

$\psi_{obj} = 0$ führt zu $\gamma_j = 1$, womit der Faktor aus der Modellgleichung in (3.6) unbedeutend wäre und das Rasch-Modell somit unverletzt. $\psi_{obj} = 0.5$ entspricht hingegen einer

[3] Der resultierende Erwartungswert ist $\exp\{0\} = 1$.

starken Verletzung (vgl. Suárez-Falcón und Glas 2003: 134). Durch die verstärkte Streuung des Diskriminationsparameters werden einige Fragen trennschärfer und andere weniger trennscharf.

Aufgrund der Verknüpfung der Diskriminationsparameter mit den Aufgabenschwierigkeiten wird angenommen, dass sich die Schätzabweichungen der Schwierigkeitsparameter tendenziell vergrößern. Die Schätzungen der Personenfähigkeiten sollten davon unbetroffen sein, da sich die Fähigkeitsparameter aus den Antworten vieler Items zusammensetzen, die sich im Mittel ausgleichen.

3.2.2 Verletzung der Eindimensionalität

Die Eindimensionalität wird dann verletzt, wenn in einem Test mehr als nur ein Konstrukt gemessen wird. Um die Konsequenzen von Mehrdimensionalität zu messen, wird auf ein Vorschlag von Glas (1992) zurückgegriffen, der auch in den Studien von Suárez-Falcón und Glas sowie Mair und Ledl verwendet wird. Demnach ist der Personenparameter θ_i kein eindimensionales Konstrukt. Stattdessen werden zwei latente Fähigkeiten erzeugt, die miteinander in Verbindung stehen:

$$\begin{pmatrix} \theta_i^A \\ \theta_i^B \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{\theta^A\theta^B} \\ \sigma_{\theta^A\theta^B} & 1 \end{pmatrix} \right) \tag{3.9}$$

Die beiden Personenfähigkeiten θ_i^A und θ_i^B werden bivariat standardnormalverteilt erzeugt. Zwischen ihnen besteht eine Kovarianz (welche hier äquivalent zur Korrelation ist) von $\sigma_{\theta^A\theta^B}$. Als Ausmaß der Annahmeverletzung der Eindimensionalität heißt die Kovarianz hier auch ψ_{dim}. Sie wird als nichtnegativ angenommen und in jedem Simulationsdurchlauf zufällig aus einer stetigen Gleichverteilung zwischen 0 und 1 bestimmt:

$$\psi_{dim} = \sigma_{\theta^A\theta^B} \sim \mathcal{U}_S[0,1] \tag{3.10}$$

$\psi_{dim} = 1$ ist gleichbedeutend damit, dass es sich bei θ_i^A und θ_i^B um die identische Fähigkeit einer Person handelt. θ_i wäre also eindimensional und die Modellannahme nicht verletzt. Bei $\psi_{dim} = 0$ besteht zwischen den beiden latenten Variablen überhaupt kein Zusammenhang.

Es wird davon ausgegangen, dass jede Frage entweder nur auf θ_i^A oder nur auf θ_i^B wirkt. Diese Annahme wird auch *between-item multidimensionality* genannt (vgl. Strobl 2012: 62). Dementsprechend gibt es nun zwei Wahrscheinlichkeitsfunktionen zur Erzeugung einer Datenmatrix. Für die Fragen j_A und die dazugehörigen θ_i^A gilt:

$$\mathcal{P}(X_{i,j_A} = 1 | \theta_i^A, \delta_{j_A}) = \frac{\exp(\theta_i^A - \delta_{j_A})}{1 + \exp(\theta_i^A - \delta_{j_A})} \tag{3.11}$$

Daneben gilt $\mathcal{P}(X_{i,j_B} = 1 | \theta_i^B, \delta_{j_B})$ analog.

In der Monte-Carlo-Simulation wird für jede Frage eine der beiden Gleichungen (und damit eine der beiden latenten Variablen θ_i^A oder θ_i^B) zufällig bestimmt. Beide haben also die gleiche Auswahlwahrscheinlichkeit.

Wenn der Zusammenhang der beiden Zielvariablen schwächer wird (oder überspitzt, wenn statt einer nun zwei latente Variablen zugrunde liegen), sollten sich hypothetisch vor allem die Schätzabweichungen der Fähigkeitsparameter vergrößern. Denn mit dem Rasch-Modell wird lediglich ein Konstrukt erfasst.

Es ist zu erwarten, dass sich die Schätzabweichungen der Schwierigkeitsparameter ebenfalls eher vergrößern. Denn die Aufgaben sind jeweils nur einem der beiden Faktoren zugeordnet und die Schwierigkeiten der Aufgaben bestimmen sich dadurch, wie viele Befragte diese (nicht) beantworten konnten.

3.2.3 Verletzung der lokalen stochastischen Unabhängigkeit

Lokale stochastische Abhängigkeit besteht, wenn die Beantwortung der einen Frage von einer anderen abhängt. Jannarone (1986) hat zu diesem Zweck ein Modell vorgeschlagen, welches das Rasch-Modell erweitert:

$$\mathcal{P}(X_{i,j} = 1 | X_{i,k} = x_{i,k}; \theta_i, \delta_j, \psi_{dep}) = \frac{\exp(\theta_i - \delta_j + \psi_{dep} x_{i,k})}{1 + \exp(\theta_i - \delta_j + \psi_{dep} x_{i,k})} \tag{3.12}$$

$X_{i,j}$ und $X_{i,k}$ sind Zufallsvariablen für die Antworten auf die Fragen j und k. Dabei hängt $X_{i,j}$ von $X_{i,k}$ ab – für alle Paare der Kombinationen $(j,k) = (m, m-1), (m-2, m-3), \ldots, (2,1)$ (wobei m eine gerade Zahl ist).

Durch den Faktor ψ_{dep} wird die Stärke der lokalen Abhängigkeit ausgedrückt. Für die Monte-Carlo-Simulation wird ψ_{dep} für jedem Durchgang aus einer stetigen Gleichverteilung erzeugt:

$$\psi_{dep} \sim \mathcal{U}_S[0,1] \tag{3.13}$$

$\psi_{dep} = 0$ entspricht stochastischer Unabhängigkeit, während $\psi_{dep} = 1$ eine starke Abhängigkeit darstellt.

Hypothetisch sollte eine zunehmende Verletzung der Annahme die Schätzabweichungen sowohl von Personen- als auch von Aufgabenparametern verstärken. Durch den hinzugekommenen Abhängigkeitsfaktor sinkt die Bedeutung der (eigentlichen) Schwierigkeit δ_j für die Lösung einer Aufgabe – falls diese von einer vorhergehenden Antwort abhängt.[4] Insofern sind davon auch die Schätzungen der Fähigkeitsparameter betroffen, deren Abweichungen mit wachsendem ψ_{dep} steigen müssten.

Wang und Wilson (2005) ermitteln große Schätzabweichungen und Verzerrungen im Zusammenhang mit lokaler stochastischer Abhängigkeit und schlagen deshalb ein Rasch-Testlet-Modell vor.

[4] Der Schwierigkeitsparameter δ_j wird als unabhängig von ψ_{dep} angenommen.

3.3 Generierung der Datenmatrix

Im Unterschied zu Suárez-Falcón und Glas sollen die einzelnen Verletzungsszenarien nicht getrennt, sondern simultan betrachtet werden. Dazu ist es erforderlich alle Verletzungen in eine gemeinsame Modellgleichung zu implementieren, aus der die Lösungswahrscheinlichkeiten für alle Aufgaben und Personen generiert werden. Aus den Gleichungen (3.6), (3.11) und (3.12) wird:

$$\pi_{i,j} = \mathcal{P}(X_{i,j} = 1 | X_{i,k} = x_{i,k}; \theta_i^*, \delta_j, \gamma_j, \psi_{dep}) = \frac{\exp\{\gamma_j(\theta_i^* - \delta_j + \psi_{dep} x_{i,k})\}}{1 + \exp\{\gamma_j(\theta_i^* - \delta_j + \psi_{dep} x_{i,k})\}} \tag{3.14}$$

mit

$$\theta_i^* = \begin{cases} \theta_i^A & \forall j \in m_A \\ \theta_i^B & \forall j \in m_B \end{cases} \tag{3.15}$$

sowie allen in diesem Kapitel aufgeführten Eigenschaften. Die spezifischen Parameter für die Modellverletzungen (3.8), (3.10) und (3.13) sind unabhängig voneinander.

Die Datenmatrix wird mithilfe der einzelnen $\pi_{i,j}$ für alle n Befragten und alle m Fragen erzeugt. Die Antworten auf die einzelnen Aufgaben werden bernoulliverteilt erzeugt:

$$f(x_{i,j}) = \pi_{i,j}^{x_{i,j}} (1 - \pi_{i,j})^{1 - x_{i,j}} \tag{3.16}$$

3.4 Schätzung des Rasch-Modells

Zur Schätzung des Rasch-Modells wird das R-Paket `eRm` (kurz für *Extended Rasch Modeling*) von Mair et al. (2012) genutzt. Diese nutzt das bedingte Maximum-Likelihood-Verfahren, das in Abschnitt 2.3 ausführlich beschrieben wurde.

4 Statistische Methoden

4.1 Abweichungsmaße

Die durch das Rasch-Modell geschätzten Parameter $\hat{\theta}_i$ und $\hat{\delta}_j$ sollen mit den „echten“ (generierten) Werten θ_i und δ_j verglichen werden. Für diesen Zweck werden zwei Maße vorgestellt, die unterschiedliche Dimensionen von Schätzabweichungen quantifizieren.

4.1.1 Mittlere betragsmäßige Abweichung

Die mittlere betragsmäßige Abweichung (*mean absolute deviation*, *MAD*) beschreibt den Durchschnittsfehler eines Schätzers vom wahren Wert – ohne Beachtung des Vorzeichens.

Bei Betrachtung der Aufgabenparameter δ_j und der dazugehörigen Schätzwerte $\hat{\delta}_j$ gilt:

$$MAD_{\hat{\delta}} = \frac{1}{m} \sum_{j=1}^{m} |\hat{\delta}_j - \delta_j| \tag{4.1}$$

Die mittlere betragsmäßige Schätzabweichung der Personenparameter ist (technisch betrachtet) der Durchschnitt zweier mittlerer betragsmäßiger Abweichungen: die Abweichungen der Schätzwerte $\hat{\theta}_i$ von den wahren Fähigkeiten im Konstrukt A θ_i^A sowie die Abweichungen der Schätzwerte $\hat{\theta}_i$ von den wahren Fähigkeiten im Konstrukt B θ_i^B. Damit gilt:

$$\begin{aligned} MAD_{\hat{\theta}} &= \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i^A| + \frac{1}{n} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i^B| \right) \\ &= \frac{1}{2n} \sum_{i=1}^{n} \left(|\hat{\theta}_i - \theta_i^A| + |\hat{\theta}_i - \theta_i^B| \right) \end{aligned} \tag{4.2}$$

Die beiden wahren Personenparameter θ_i^A und θ_i^B wurden in Abschnitt 3.2.2 eingeführt. Damit wird die Verletzung der Annahme der Eindimensionalität generiert, deren Einfluss auf Schätzabweichungen untersucht wird. Im Falle gleicher Personenparameter ($\theta_i^A = \theta_i^B$) entspricht 4.2 technisch 4.1.

4.1.2 Bias

Während die mittlere betragsmäßige Abweichung das durchschnittliche Ausmaß einer Schätzabweichung quantifiziert, beschreibt der Bias eine systematische Über- oder Unterschätzung eines Schätzers vom wahren Wert. Entspricht der Schätzer im Mittel dem wahren

Wert, heißt er erwartungstreu. In einer Stichprobe sollte er möglichst nahe am wahren Wert liegen.

Würde man die Simulationsparameter auf den einfachen Bias für die Aufgabenparameter regressieren, erhielte man einen bedingten Erwartungswert des Bias. Problematisch ist aber, dass sich Über- und Unterschätzungen in der Regression ausgleichen (können) und die Simulationsparameter in diesem Fall scheinbar keinen Einfluss auf den Bias hätten. Um dies zu umgehen, wird der betragsmäßige Bias (*absolute bias*, *AB*) verwendet.

Für die Aufgabenparameter δ_j und die entsprechenden Schätzungen $\hat{\delta}_j$ gilt:

$$AB_{\hat{\delta}} = \frac{1}{m} | \sum_{j=1}^{m} \left(\hat{\delta}_j - \delta_j \right) | \tag{4.3}$$

Der Bias der Fähigkeitenschätzungen ist der Mittelwert des betragsmäßigen Bias der Schätzungen $\hat{\theta}_i$ θ_i^A und des betragsmäßigen Bias von $\hat{\theta}_i$ zu θ_i^A:

$$\begin{aligned} AB_{\hat{\theta}} &= \frac{1}{2} \left(| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i^A \right) | + | \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i^B \right) | \right) \\ &= \frac{1}{2n} \left(| \sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i^A \right) | + | \sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i^B \right) | \right) \end{aligned} \tag{4.4}$$

4.2 Regressionsanalyse

Anders als in der Referenzstudie von Suarez-Falcon und Glas werden die Simulationsparameter n (Anzahl der Befragten), m (Anzahl der Fragen), ψ_{obj} (die Standardabweichung des Diskriminationsparameters γ für die Verletzung der spezifischen Objektivität), ψ_{dim} (die Korrelation der Personenfähigkeitsparameter θ^A und θ^B für Multidimensionalität) und ψ_{dep} (dem Faktor der lokalen stochastischen Abhängigkeit) nicht deterministisch festgelegt, sondern zufällig aus vorgegebenen Wertebereichen gezogen.

Mithilfe eines Regressionsmodells ergibt sich die Möglichkeit die Simulationsparameter auf die Abweichungsmaße zu regressieren und damit statistisch zu prüfen, welche Parameter den größten Einfluss auf Schätzabweichungen haben.[1]

4.2.1 Lineare Regression

Das (multiple) lineare Regressionsmodell wird verwendet, um lineare Zusammenhänge zwischen einer metrisch-skalierten abhängigen Variablen y und einer oder mehreren unabhängigen Variablen X zu untersuchen. Die abhängige Variable ist die Schätzabweichung, die mit einem der in Abschnit 4.1 vorgestellten Abweichungsmaße quantifiziert wird. Die

[1] Zum weiteren Verständnis, den Annahmen und zur Schätzung der Regressionskoeffizienten im linearen Regressionsmodell siehe etwa Greene (2011: Kap. 2–4) oder Fahrmeir et al. (2007: Kap. 2–3).

unabhängigen Variablen sind die Simulationsparameter. Die Modellgleichung lautet:

$$y_r = \beta_0 + \beta_1 X_{r,1} + \ldots + \beta_p X_{r,p} + \varepsilon_r \tag{4.5}$$

$\beta_0, \beta_1, \ldots, \beta_p$ sind die $p+1$ unbekannten Regressionskoeffizienten. Dabei symbolisiert β_0 die Regressionskonstante (bzw. den Achsenabschnitt auf der y-Achse), während die übrigen β-Koeffizienten Regressionsgewichte der unabhängigen Variablen X darstellen. ε ist eine Störgröße mit dem Mittelwert 0. Der Index r markiert den Simulationsdurchlauf ($r = 1, \ldots, 10.000$), der eine Beobachtung im Regressionsmodell repräsentiert.

Die unbekannten Regressionskoeffizienten können im einfachsten Fall mithilfe der Methode der kleinsten Quadrate (englisch: *ordinary least squares*, kurz OLS) geschätzt werden. Der Schätzwert $\hat{\beta}$ (marginaler Effekt) gibt an, wie sich die abhängige Variable verändert, wenn sich die betrachtete unabhängige Variable um eine Einheit erhöht (*ceteris paribus*: bei Konstanthaltung aller anderen Einflussvariablen).

Da der Wertebereich durch die betragsmäßigen Abweichungen positiv beschränkt ist, ist es denkbar, dass das klassische lineare Regressionsmodell zu inkonsistenten Schätzungen führt (vgl. Amemya 1973). Ein Regressionsmodell, das sich zur Analyse gestutzter abhängiger Variablen eignet, ist das Tobit-Modell. Allerdings führen die Regressionsschätzungen zu sehr ähnlichen Resultaten (identisch bis zur 5. Nachkommastelle der Regressionskoeffizienten), sodass auf die Beschreibung und Anwendung des Tobit-Modells verzichtet wird.

4.2.2 Robuste Regression

Die klassische OLS-Regression ist nicht robust gegenüber Ausreißern. Für erwartungstreue Regressionsschätzungen sollten die Daten auf Ausreißer geprüft und eventuell um diese bereinigt werden. Ein erstes Hilfsmittel dafür ist ein Boxplot. Vor allem beim betragsmäßigen Bias der Aufgabenparameter fallen zwei Extremwerte auf (Abbildung 4.1 links). Während nahezu alle Werte sehr nahe bei 0 liegen, lauten die Extremwerte 0,0156 und 0,0015. Durch Ausschluss der beiden Werte sinkt AB_δ von 1,71 E-06 auf 2,15 E-17. Im bereinigtem Boxplot (rechts) lässt sich die AB_δ-Verteilung anschließend sehr viel besser erkennen.[2] Augenfällig ist aber auch, dass weitere Ausreißer verbleiben. Allerdings liefern unterschiedliche Ausreißertests (etwa der Grubbs-Test und der Walsh-Test) keine einheitlichen Aussagen darüber, wie viele weitere Werte ausgeschlossen werden sollten.[3]

Eine Lösung für dieses Problem bieten robuste Verfahren, sogenannte M-Schätzer, welche großen Abweichungen $e_r = \hat{y}_r - y_r$ kleinere (oder gar keine) Gewichte in der Regression zuweisen. Während der Huber-k-Schätzer jedem Wert ein Gewicht > 0 gibt, berücksichtigen etwa der Hampel-Schätzer oder Tukey's biweight sehr starke Ausreißer gar nicht mehr. Für eine allgemeine Übersicht über die wichtigsten M-Schätzer und deren Algorithmen siehe

[2] Neben $AB\delta$ ist mit MAD_δ auch das andere Abweichungsmaß für die Schätzungen der Aufgabenparameter von den beiden außergewöhnlichen Ausreißern betroffen. Da die Personenparameter im Rasch-Modell mithilfe der geschätzten Aufgabenparameter ermittelt werden, ist es sinnvoll alle Werte der beiden Simulationsdurchläufe auszuschließen.

[3] Weil die Verwendung von Grubbs- und Walsh-Test letzlich ohne Konsequenzen bleibt, wird auf deren Vorstellung hier verzichtet.

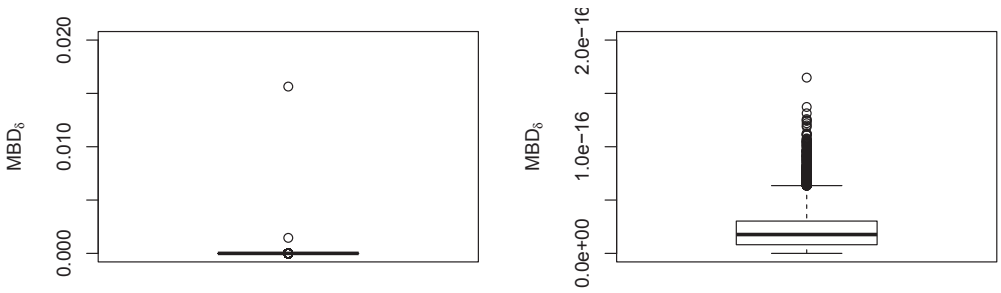

Abbildung 4.1: Boxplots von AB_δ ohne (links) bzw. nach (rechts) Ausschluss zweier Ausreißer

etwa Rönz (2001: 49–60).

Für die robuste Regression wird die Gewichtsfunktion von Tukey's biweight bevorzugt, weil sie den (bedingten) Median relativ zu den anderen Werten am stärksten gewichtet. Dessen Gewichtsfunktion $w(z)$ lautet:

$$w(z) = \begin{cases} \left(1 - \frac{z_r^2}{a^2}\right)^2 & |z_r| \leq a \\ 0 & |z_r| > a \end{cases} \tag{4.6}$$

Dabei ist a eine Konstante, die auf 4,685 fixiert ist[4], und z_r die Standardisierung von e_r:

$$z_r = \frac{e_r}{\text{median absolute deviation}} = \frac{\hat{y}_r - y_r}{\text{median absolute deviation}} \tag{4.7}$$

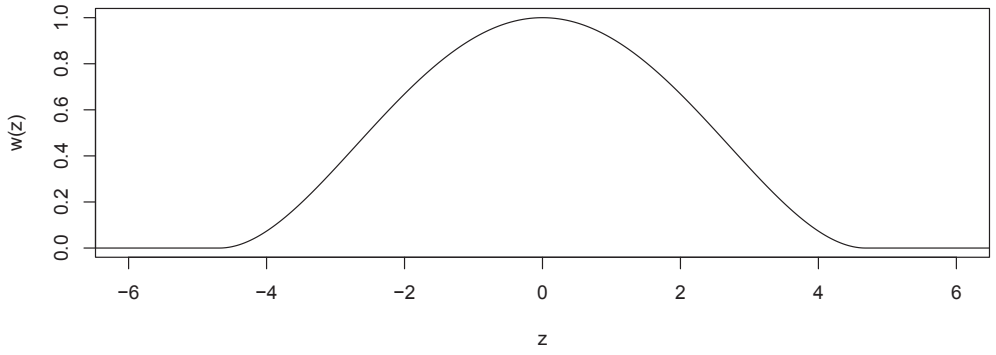

Abbildung 4.2: Gewichtsfunktion $w(z)$ von Tukey's biweight ($a = 4{,}685$) in Abhängigkeit von z

[4] Sowohl in SPSS als auch im verwendeten R-Paket `MASS` ist 4,685 der Standardwert für a. Je größer a gewählt wird, desto mehr Werte werden tendenziell berücksichtigt und desto größer werden die Gewichte $w(z)$.

Übersteigt $|z_r|$ also 4,685, wird die Beobachtung in der Regression nicht berücksichtigt. Andernfalls erhält es ein Gewicht, das umso größer wird, je kleiner $|z_r|$ (und damit $|e|$) ist. Nur bei $|z_r| = 0$ geht die Beobachtung mit vollem Gewicht $w(z) = 1$ in die Regression ein. Abbildung 4.2 veranschaulicht diese Zusammenhänge.

Die eigentliche Schätzung des Regressionsmodell ist ein iteratives Verfahren, das mithilfe des Maximum-Likelihood-Ansatzes erfolgt (vgl. etwa Rousseeuw und Leroy 2003).

4.2.3 Bootstrapping

Die generierten Daten, die in der Regression verwendet werden, stellen eine Zufallsstichprobe S dar, deren wahre Verteilung F unbekannt ist. Signifikanztests sind deshalb unzulässig, weil sie eine bekannte asymptotische Verteilung der Parameter voraussetzen. Durch wiederholtes Ziehen mit Zurücklegen kann aus dieser Zufallsstichprobe S eine empirische Verteilung $\hat{F}$ erzeugt werden, mit der inferenzstatistische Aussagen getroffen werden können. Dieses Resamplingverfahren ist auch als Bootstrapping bekannt (vgl. Efron 1979).

Für eine lineare Regression kommen grundsätzlich mehrere Bootstrap-Verfahren infrage. In dieser Arbeit wird der Bootstrap der Beobachtungen (auch als *case resampling* oder Vektor-Sampling bezeichnet) angewendet, wobei die unabhängigen Variablen als zufällig betrachtet werden (im Gegensatz zum Bootstrap der Residuen, vgl. Fox 2008: 597). Dabei werden alle im Regressionsmodell verwendeten Daten in einer Matrix zusammengefasst:

$$S_{9.998,p+1} = \begin{pmatrix} y_1 & X_{1,1} & \cdots & X_{1,p} \\ y_2 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & \vdots & & \vdots \\ y_{9.998} & X_{9.998,1} & \cdots & X_{9.998,p} \end{pmatrix} \tag{4.8}$$

Aus den 9.998 verbliebenen Simulationsdurchläufen (wobei jeder Durchlauf als Zeilenvektor betrachtet wird) werden 2.000 Bootstrap-Stichproben (mit Zurücklegen) gezogen. Aus der „neuen“ Datenmatrix (die also 2.000 Zeilen und somit 2.000 Beobachtungen umfasst) wird das Regressionsmodell erneut gerechnet. Die geschätzten Regressionskoeffizienten $\hat{\beta}_0^{boot}, \hat{\beta}_1^{boot}, \ldots, \hat{\beta}_p^{boot}$ werden gespeichert. Aus der ursprünglichen Datenmatrix werden analog noch 1.999 weitere Datenmatrizen erzeugt und Regressionen gerechnet, woraus für jeden Regressionskoeffizienten 2.000 Schätzungen resultieren, die eine empirische Verteilungsfunktion bilden. Die Mittelwerte der Regressionskoeffizienten aus den 2.000 Durchläufen $\bar{\hat{\beta}}_0^{boot}, \bar{\hat{\beta}}_1^{boot}, \ldots, \bar{\hat{\beta}}_p^{boot}$ sind schließlich die Bootstrap-korrigierten Schätzwerte der Regressionskoeffizienten.

Mithilfe der empirischen Verteilung der Bootstrap-Schätzungen lassen sich Konfidenzintervalle für die Koeffizienten berechnen, welche unabhängig von den Standardfehlern erzeugt werden. Das sogenannte Bootstrap-Perzentil-Intervall ordnet die jeweils 2.000 Bootstrap-Schätzwerte – für jedes $\beta_0, \beta_1, \ldots, \beta_p$ unabhängig – der Größe nach und zieht (beim Signifikanzniveau $\alpha = 0{,}05$) den 50. Wert als untere Schranke und den 1.950. Wert als obere Schranke des Konfidenzintervalls. Wird der Wert 0 vom Konfidenzintervall überdeckt, gilt der Regressionskoeffizient als insignifikant.

5 Ergebnisse

In diesem Kapitel werden die Simulationsergebnisse für die Schätzabweichungen der Personen- und Aufgabenparameter im Rasch-Modell präsentiert und bewertet. Die Simulationsparameter, welche in Kapitel 3 beschrieben wurden, sind in Tabelle 5.1 zusammengefasst.[1][2] Für die Abweichungen sollte berücksichtigt werden, dass die einzelnen (generierten und geschätzten) Personen- und Aufgabenparameter standardnormalverteilt sind.

Fähigkeit einer Person i	θ_i	$\sim \mathcal{N}(0,1)$
Schwierigkeit einer Aufgabe j	δ_j	$\sim \mathcal{N}(0,1)$
Anzahl an Personen	n	$\sim \mathcal{U}_D[100,1.000]$
Anzahl an Aufgaben	m	$\sim 2 \cdot \mathcal{U}_D[5,50]$
Verletzung der spezifischen Objektivität	ψ_{obj}	$\sim \mathcal{U}_S[0,0{,}5]$
Verletzung der Eindimensionalität	ψ_{dim}	$\sim \mathcal{U}_S[0,1]$
Verletzung der lokalen stochastischen Unabhängigkeit	ψ_{dep}	$\sim \mathcal{U}_S[0,1]$

Tabelle 5.1: Übersicht der Simulationsparameter für jeden der 9.998 Durchläufe

Um den Einfluss der einzelnen Parameter auf die Schätzabweichungen beurteilen zu können, wird eine lineare Regressionsanalyse durchgeführt. Darin stellt jeder der 9.998 Simulationsdurchläufe eine Beobachtung dar.[3] n, m, ψ_{obj}, ψ_{dim}[4] und ψ_{dep} sind die unabhängigen Variablen, welche auf die in Abschnitt 4.1 vorgestellten Abweichungsmaße MAD und AB (jeweils für θ und δ) regressiert werden. Es wird ein Schätzverfahren verwendet, das robust gegenüber Ausreißern ist (vgl. Abschnitt 4.2.2) und darüber hinaus durch Bootstrapping (vgl. Abschnitt 4.2.3) inferenzstatistische Aussagen erlaubt. Für die Beurteilung eines marginalen Effektes wird $\bar{\hat{\beta}}^{boot}$ (Durchschnitt der Boostrap-Schätzwerte von 2.000 durchgeführten robusten Regressionen) herangezogen. Dabei werden die anderen Einflussvariablen als konstant angenommen. Sind die Annahmen unverletzt, werden die ψ-Koeffizienten für die Berechnung der bedingten Erwartungswerte ignoriert.

[1] Die Aufgabenschwierigkeiten δ_j werden im Anschluss der Generierung zentriert.

[2] Relevante statistische Kennzahlen sind im Anhang aufgeführt.

[3] Zwei Durchläufe wurden aufgrund ihrer Extremwerte von der weiteren Analyse ausgeschlossen (siehe Abschnitt 4.2.2).

[4] Zur Erleichterung der Interpretation wurde der transformierte Parameter $1 - \psi_{dim}$ in die Regression aufgenommen, da die Annahme der Eindimensionalität bei $\psi_{dim} = 1$ unverletzt ist und $\psi_{dim} = 0$ den maximal angenommenen Verletzungsgrad ausdrückt.

5.1 Schätzabweichungen der Personenparameter

Im Idealfall soll das Rasch-Modell die echten Fähigkeiten von Personen möglichst präzise schätzen. Je weiter die Schätzungen von der (hypothetischen) Wirklichkeit abweichen, desto unzuverlässiger wird das Testinstrument.

Mittlere betragsmäßige Abweichung

Um den möglichen Einfluss von Annahmeverletzungen auf die mittlere betragsmäßige Abweichung der Fähigkeitsparameter einzuschätzen, werden zunächst Streudiagramme betrachtet (siehe Abbildung 5.1[5]). Dabei scheint es, dass die Verletzung der spezifischen Objektivität keinen Effekt auf $MAD_{\hat{\theta}}$ hat, Multidimensionalität und lokale stochastische Abhängigkeit hingegen schon.

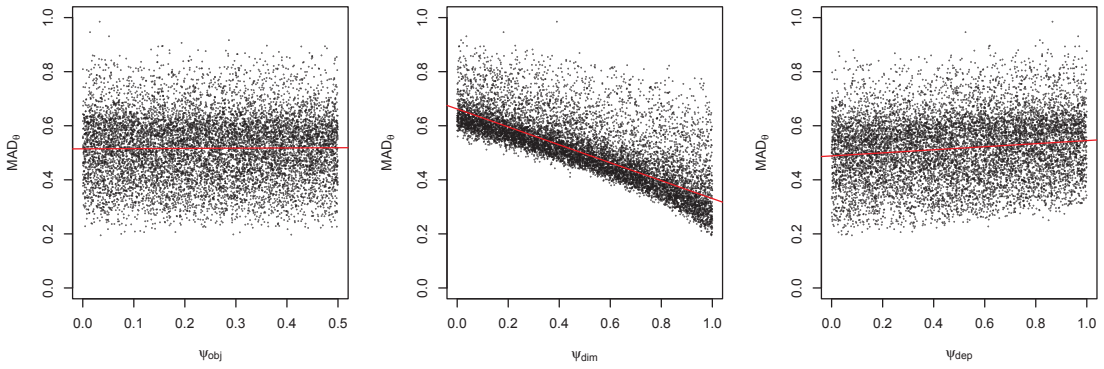

Abbildung 5.1: $MAD_{\hat{\theta}}$ (y-Achse) in Abhängigkeit von ψ_{obj} (links), ψ_{dim} (Mitte) und ψ_{dep} (rechts), rot: isolierte Regressionsgeraden

	$\bar{\hat{\beta}}^{boot}$	SE($\bar{\hat{\beta}}^{boot}$)	$\hat{\beta}^{boot}_{2,5\%}$	$\hat{\beta}^{boot}_{97,5\%}$
Konstante	0,420341	0,002593	0,415395	0,425531
n	-0,000003	0,000001	-0,000006	-0,000001
m	-0,001894	0,000029	-0,001951	-0,001835
ψ_{obj}	0,000661	0,002170	-0,003368	0,004996
$1-\psi_{dim}$	0,325142	0,001302	0,322616	0,327730
ψ_{dep}	0,056282	0,001148	0,054091	0,058673

Tabelle 5.2: Kennwerte der robusten Regression mit Bootstrap für $MAD_{\hat{\theta}}$

Tabelle 5.2 zeigt die Kennwerte der robusten linearen Regression auf die mittlere betragsmäßige Abweichung der Fähigkeitsparameter. Ohne Modellverletzungen liegt die mittlere

[5] Die roten Geraden geben die Regressionsgeraden für die isolierten Modellverletzungen an. Sie sind das Ergebnis einer robusten Regression ohne Boostrapping.

betragsmäßige Abweichung bei 10 absolvierten Aufgaben und 100 Befragten (dem Minimum, das in der Simulation angenommen wird) bei rund 0,4009. Bei 100 absolvierten Aufgaben und 100 Teilnehmern verringert sie sich auf etwa 0,2308. Der Regression zufolge wirkt sich die Anzahl an Aufgaben maßgeblich auf die Schätzgenauigkeit aus. Die Anzahl der Befragten n ist hingegen schwach wirksam. Jeder zusätztliche Testteilnehmer verringert $MAD_{\hat{\theta}}$ um schätzungsweise 0,000003. Bei 100 Aufgaben und 1.000 Teilnehmern (dem Maximum an Aufgaben und Befragten) bedeutet dies eine mittlere betragsmäßige Abweichung von ca. 0,2281.

Der Eindruck, der aus den Streudiagrammen in Abbildung 5.2 entstanden ist, bestätigt sich in der Regressionsschätzung. Die Verletzung der spezifischen Objektivität, deren Schweregrad mit ψ_{obj} ausgedrückt wird, hat keinen signifikanten Einfluss. Hochsignifikant wirken sich allerdings Multidimensionalität und lokale stochastische Abhängigkeit aus. Gibt es keine Korrelation zwischen zwei zugrunde liegenden latenten Variablen θ^A und θ^B (also $1 - \psi_{dim} = 1$), vergrößert sich die Schätzabweichung um durchschnittlich 0,3251. Nicht ganz so drastisch, aber dennoch bedeutsam wirkt sich die lokale stochastische Abhängigkeit auf die mittlere Abweichung der Personenparameter aus. Bei maximal angenommener Abhängigkeit von $\psi_{dep} = 1$ steigt $MAD_{\hat{\theta}}$ um 0,0563.

Bias

Betrachtet man die Streudiagramme der Annahmeverletzungen im Hinblick auf den durchschnittlichen Bias der Personenparameter (Abbildung 5.2), fällt vor allem ein starker Zusammenhang mit der lokalen stochastischen Abhängigkeit auf (rechts). Die übrigen Verletzungen scheinen keine gravierenden Auswirkungen auf $AB_{\hat{\theta}}$ zu haben.

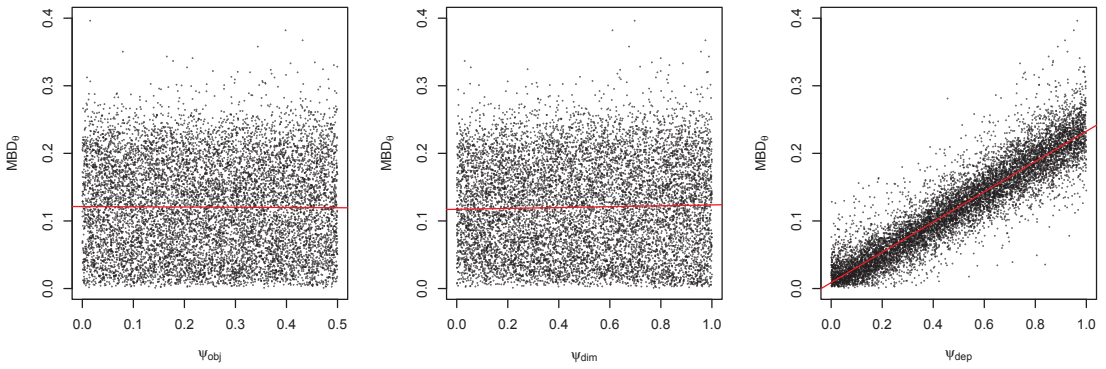

Abbildung 5.2: $AB_{\hat{\theta}}$ (y-Achse) in Abhängigkeit von ψ_{obj} (links), ψ_{dim} (Mitte) und ψ_{dep} (rechts), rot: isolierte Regressionsgeraden

Die Beobachtungen aus den Streudiagrammen werden mit dem Regressionsmodell untersucht (Tabelle 5.3). Ohne Annahmeverletzungen weist das Rasch-Modell mit 10 Aufgaben und 100 Befragten einen Bias von rund 0,016 auf, mit 100 Aufgaben und 1.000 Befragten sinkt er auf 0,007. Sowohl die Anzahl der Fragen als auch die der Personen wirken sich bedeutsam auf den Bias der Fähigkeitsparameter aus. Sind die Annahmen unverletzt, misst das Modell die Personenfähigkeiten prinzipiell erwartungstreu.

	$\bar{\hat{\beta}}^{boot}$	SE($\bar{\hat{\beta}}^{boot}$)	$\hat{\beta}^{boot}_{2,5\%}$	$\hat{\beta}^{boot}_{97,5\%}$
Konstante	0,017333	0,001124	0,015125	0,019447
n	-0,000003	0,000001	-0,000005	-0,000001
m	-0,000071	0,000010	-0,000090	-0,000051
ψ_{obj}	0,002613	0,001789	-0,000874	0,006003
$1-\psi_{dim}$	-0,006493	0,000846	-0,008262	-0,004885
ψ_{dep}	0,224501	0,000838	0,222835	0,226123

Tabelle 5.3: Kennwerte der robusten Regression mit Bootstrap für $AB_{\hat{\theta}}$

Im Regressionsmodell bestätigt sich der vermutete Einfluss von ψ_{dep}. Bei einem maximalen Abhängigkeitsfaktor von 1 steigt der Bias um 0,2245. Bei unterstellter Linearität sowie 10 Fragen und 100 Befragten würde schon ein geringer Faktor von $\psi_{dep} = 0,1$ den Bias mehr als verdoppeln (auf 0,0387).

Interessanterweise sinkt der Bias der Fähigkeitenparameter mit steigender Multidimensionalität. Sind θ^A und θ^B völlig unkorreliert ($1-\psi_{dim} = 1$), verringert sich $AB_{\hat{\theta}}$ um schätzungsweise 0,0065. Warum steigende Multidimensionalität zu einem kleineren Bias führt, ist allerdings unklar.

Keinen signifikanten Einfluss hat die Verletzung der spezifischen Objektivität. Damit wirkt sie sich insgesamt nicht bedeutsam auf die Schätzergebnisse der Personenparameter aus.

5.2 Schätzabweichungen der Aufgabenparameter

Das Rasch-Modell besitzt den Vorteil adaptives Testen zu ermöglichen, indem getestete Items für weitere Tests genutzt werden können (mit neuen Befragten oder zur Wiederholungsmessung). Weichen die geschätzten Augabenschwierigkeiten jedoch stark von den wahren Schwierigkeiten ab oder sind sie sogar verzerrt, lassen sich die Items nicht ohne Vorbehalt für weitere Tests nutzen.

Mittlere betragsmäßige Abweichung

Bei Betrachtung der Streudiagramme der Annahmeverletzungen im Zusammenhang mit der mittleren betragsmäßigen Abweichung der Aufgabenparameter (Abbildung 5.3), wird erneut ein starker Einfluss der lokalen stochastischen Abhängigkeit deutlich (rechts). Darüber hinaus scheint $MAD_{\hat{\delta}}$ auch mit zunehmenden Verletzungsgrad der spezifischen Objektivität zu steigen (links).

In der Regression auf $MAD_{\hat{\delta}}$ wirken alle Schätzkoeffizienten bedeutsam (Tabelle 5.4). Offensichtlich führt eine Erhöhung der Aufgaben zu einer Vergrößerung der mittleren betragsmäßigen Abweichung. Ohne Annahmeverletzungen und mit 100 Testpersonen liegt sie zwischen 0,0905 (10 Aufgaben) und 0,0967 (100 Aufgaben). 1.000 Befragte sorgen allerdings dafür, dass die durchschnittliche Abweichung (bei 10 Aufgaben) auf minimal 0,0392 sinkt. Eine zusätzliche Testperson gleicht eine erhöhte mittlere Abweichung, die durch die

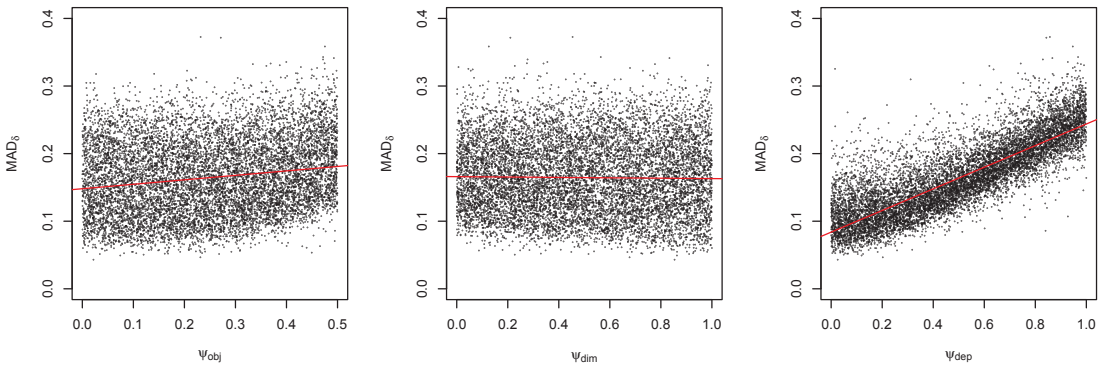

Abbildung 5.3: $MAD_{\hat{\delta}}$ (y-Achse) in Abhängigkeit von ψ_{obj} (links), ψ_{dim} (Mitte) und ψ_{dep} (rechts), rot: isolierte Regressionsgeraden

Zunahme einer weiteren Aufgabe entstanden ist, dem Regressionsmodell zufolge ungefähr aus.

	$\hat{\beta}^{boot}$	SE($\hat{\beta}^{boot}$)	$\hat{\beta}^{boot}_{2,5\%}$	$\hat{\beta}^{boot}_{97,5\%}$
Konstante	0,095433	0,001159	0,093305	0,097784
n	-0,000057	0,000001	-0,000059	-0,000055
m	0,000070	0,000010	0,000050	0,000088
ψ_{obj}	0,066440	0,001761	0,062938	0,069880
$1-\psi_{dim}$	0,003246	0,000811	0,001674	0,004848
ψ_{dep}	0,157135	0,000939	0,155155	0,158903

Tabelle 5.4: Kennwerte der robusten Regression mit Bootstrap für $MAD_{\hat{\delta}}$

Der Eindruck aus den Streudiagrammen, dass die lokale stochastische Abhängigkeit den größten Einfluss auf $MAD_{\hat{\delta}}$ hat, wird durch das Regressionsmodell gestärkt. Bei der höchsten angenommenen Aufgabenabhängigkeit $\psi_{dep} = 1$ verfünffacht sich $MAD_{\hat{\delta}}$ (bei 10 Aufgaben und 1.000 Befragten) von 0,0392 auf 0,1963, falls die übrigen Annahmen völlig unverletzt sind.

Tatsächlich hat auch die Verletzung der spezifischen Objektivität – die für die Personenparameter noch bedeutungslos war – eine beachtliche Wirkung auf $MAD_{\hat{\delta}}$. Bei maximal angenommenem Verletzungsgrad von $\psi_{obj} = 0{,}5$ wird die durchschnittliche Abweichung der Aufgabenparameter (bei 10 Aufgaben und 1.000 Befragten) um 0,0332 auf 0,0725 fast verdoppelt.

Obwohl Multidimensionalität im Vergleich zu den beiden aufgeführten Annahmeverletzungen die geringste Bedeutung auf $MAD_{\hat{\delta}}$ hat, wirkt auch sie sich problematisch aus. Bei völliger Unkorrliertheit von θ^A und θ^B steigt die durchschnittliche Abweichung um schätzungsweise 0,0032.

Bias

Die Verzerrungen der Aufgabenparameter haben ein sehr viel geringeres Niveau als die bisher betracheteten Maße (Abbildung 5.4). Nach Ausschluss zweier Ausreißer (siehe Abschnitt 4.2) beträgt der durchschnittliche betragsmäßige Bias im Datensatz 0,215 E-16 bzw. 0,0000000000000000215.[6]

Die Streudiagramme sind wenig aufschlussreich. Es können noch keine eindeutigen Effekte durch Modellverletzungen erkannt werden.

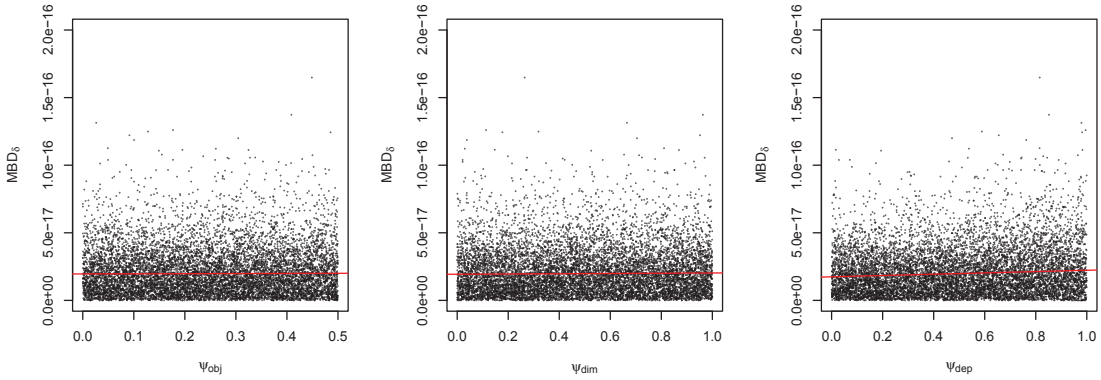

Abbildung 5.4: $AB_{\hat{\delta}}$ (y-Achse) in Abhängigkeit von ψ_{obj} (links), ψ_{dim} (Mitte) und ψ_{dep} (rechts), rot: isolierte Regressionsgeraden

	$\bar{\hat{\beta}}^{boot}$	SE($\bar{\hat{\beta}}^{boot}$)	$\hat{\beta}^{boot}_{2,5\%}$	$\hat{\beta}^{boot}_{97,5\%}$
Konstante	0,173367 E-16	0,007041 E-16	0,160345 E-16	0,187090 E-16
n	0,000004 E-16	0,000006 E-16	-0,000008 E-16	0,000016 E-16
m	0,000014 E-16	0,000061 E-16	-0,000108 E-16	0,000131 E-16
ψ_{obj}	0,010954 E-16	0,011351 E-16	-0,011861 E-16	0,033997 E-16
$1-\psi_{dim}$	-0,009576 E-16	0,005303 E-16	-0,020290 E-16	0,000817 E-16
ψ_{dep}	0,048924 E-16	0,005720 E-16	0,037610 E-16	0,059538 E-16

Tabelle 5.5: Kennwerte der robusten Regression mit Bootstrap für $AB_{\hat{\delta}}$

Die Regressionsergebnisse legen nahe, dass sich – im Gegensatz zu den Personenparametern – weder die Anzahl der Items noch die der Testteilnehmer verzerrend auf die Aufgabenparameter auswirken (Tabelle 5.5). Bei Modellgültigkeit liegt der Erwartungswert des betragsmäßigen Bias bei 0,173 E-16.[7]

Die Verletzungen der spezifischen Objektivität und der Eindimensionalität sind insignifikant.

[6] Um die Lesbarkeit zu verbessern und einen relativen Vergleich zu den vorhergehenden Analysen zu erleichtern, wird im Folgenden die wissenschaftliche Notation (ENG-Format) mit E-16 verwendet, was dem Faktor 10^{-16} entspricht.

[7] n und m bleiben hier unberücksichtigt, weil das Konfidenzintervall die 0 überdeckt und die Nullhypothese, dass beide keinen Einfluss auf den Bias haben, nicht abgelehnt werden kann.

Bedeutsam ist aber die Verletzung der Annahme der lokalen stochastischen Unabhängigkeit. Bei starker Abhängigkeit ($\psi_{dep} = 1$) wird $AB_{\hat{\theta}}$ um verhältnismäßig große 0,049 E-16 erhöht.

5.3 Fazit

Im Allgemeinen sind die Schätzabweichungen der Personenfähigkeiten deutlich größer als die der Aufgabenschwierigkeiten. Während die Abweichungen bei den Personenparametern mit Zunahme weiterer Aufgaben sinken, ist es bei Aufgabenparametern umgekehrt. Hier steigen die durchschnittlichen Abweichungen mit zunehmender Anzahl an Items. Eine Steigerung der Testteilnehmer führt aber sowohl für Fähigkeiten- als auch Schwierigkeitenparameter zu geringeren mittleren betragsmäßigen Abweichungen. Bei Gültigkeit des Rasch-Modells sind die Schätzungen im Allgemeinen erwartungstreu.

Die Verletzungen der Annahmen wirken sich – wie zu erwarten war – generell negativ auf die Vorhersagegenauigkeit der Rasch-Parameter aus. Allerdings ist hier nach Personen- und Aufgabenparameter sowie nach mittlerer betragsmäßiger Abweichung und Bias zu differenzieren. Eine Übersicht der Effekte ist in Tabelle 5.6 dargestellt.[8]

	$MAD_{\hat{\theta}}$	$AB_{\hat{\theta}}$	$MAD_{\hat{\delta}}$	$AB_{\hat{\delta}}$
ψ_{obj}	●	●	●	●
$1 - \psi_{dim}$	●	●	●	●
ψ_{dep}	●	●	●	●

● Abweichung steigt unbedeutend oder sinkt ● Abweichung steigt bedeutend

Tabelle 5.6: Übersicht der Effekte von Annahmeverletzungen auf Schätzabweichungen im Rasch-Modell

Infolge der Regressionsanalysen ist die Verletzung der spezifischen Objektivität nur für die Schätzungen der Aufgabenschwierigkeiten problematisch. Vor allem die mittleren betragsmäßigen Abweichungen werden dadurch bedeutend größer. Das erschwert ein adaptives Testen, weil die Aufgaben nicht ohne weiteres auf neue Testpersonen übertragen werden können.

Die Verletzung der Annahme der Eindimensionalität wirkt sich gravierend auf die durchschnittlichen Schätzabweichungen der Fähigkeitenparameter und gering auf die der Schwierigkeitsparameter aus. Die Erwartungstreue der Schätzer bleibt von Multidimensionalität unberührt. Allerdings werden die Schätzparameter durch verschiedene Konstrukte in einem gemeinsamen Test unpräziser und deshalb vor allem Aussagen über die Fähigkeiten von Testpersonen unsicherer.

[8] Einige der Streudiagramme legen einen (schwachen) nichtlinearen Zusammenhang zwischen den Modellverletzungen und den Schätzabweichungen nahe (siehe etwa Abbildung 5.3). Aus Gründen der Modellvereinfachung und einer leichteren Interpretierbarkeit wurde auf Polynome und andere Transformationen verzichtet, welche die bedingten Erwartungswerte geringfügig verändert hätten. Die sich daraus ergebenden Aussagen (inklusive der Signifikanzen) wären aber identisch.

Lokale stochastische Abhängigkeit wird als schwerwiegendste Annahmeverletzung bewertet. Sie wirkt sich im Erwartungswert auf alle Schätzungen bedeutend negativ aus, sowohl auf die durchschnittlichen als auch auf die systematischen Abweichungen. Fragen, die aufeinander aufbauen, sind demnach unbedingt zu vermeiden.

Durch die Quantifizierung der Schätzverzerrungen aus Modellverletzungen ist es grundsätzlich möglich, eine Biaskorrektur für das Rasch-Modell zu entwickeln. Tests, mit denen sich die Verletzungen identifizieren lassen, existieren bereits (vgl. Andersen 1973, van den Wollenberg 1982 oder Glas 1988). Es besteht aber noch Forschungsbedarf, wie ein Korrekturgewicht für die einzelnen Personen und Aufgabenparameter ermittelt werden kann.

Literaturverzeichnis

Amemya, Takeshi: Regression analysis when the dependent variable is truncated normal. In: *Econometrica* 41 (1973), Nr. 6, S. 997–1016

Andersen, Erling B.: A goodness of fit test for the rasch model. In: *Psychometrika* 38 (1973), Nr. 1, S. 123–140

Birnbaum, Allan: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F. M. (Hrsg.) ; Novick, M. R. (Hrsg.): *Statistical theories of mental test scores*. Reading, Massachusetts : Addison–Wesley, 1968, S. 395–479

Carstensen, Claus H. ; Rost, Jürgen: *MULTIRA: Ein Computerprogramm für mehrdimensionale Rasch-Modelle*. 2011. – URL `http://www.multira.de`

Efron, Bradley: Bootstrap Methods: Another Look at the Jackknife. In: *The Annals of Statistics* 7 (1979), Nr. 1, S. 1–26

Fahrmeir, Ludwig ; Kneip, Thomas ; Lang, Stefan: *Regression: Modelle, Methoden und Anwendungen*. 1. Auflage. Springer, 2007

Fox, John: *Applied Regression Analysis and Generalized Linear Models*. 2. Auflage. Sage Pubn Inc, 2008

Glas, Cees A. W.: The derivation of some tests for the rasch model from the multinomial distribution. In: *Psychometrika* 53 (1988), Nr. 4, S. 525–546

Glas, Cees A. W.: A Rasch model with a multivariate distribution of ability. In: Wilson, M. (Hrsg.): *Objective measurement: Foundations, recent developments, and applications* Bd. 1. Norwood, New Jersey : Praeger, 1992, S. 236–258

Greene, William H.: *Econometric Analysis*. 7. Auflage. Prentice Hall International, 2011

Höhne, Eva ; Hölzlwimmer, Manuela: Parameterschatzung im Raschmodell. In: *Multivariate Statistik bei psychologischen Fragestellungen (Interdisziplinares Seminar)*. Institut für Statistik, Ludwig-Maximilians-Universität München, 2009

Jannarone, Robert J.: Conjunctive item response theory kernels. In: *Psychometrika* 51 (1986), Nr. 3, S. 357–373

Mair, Patrick ; Hatzinger, Reinhold ; Maier, Marco J.: *R package eRm (Extended Rasch Modeling)*, 2012

Mair, Patrick ; Ledl, Thomas: Monte Carlo Simulations for Rasch Model Tests. In: *Memorias del XX Foro Nacional de Estadistica* (2006), S. 83–94

Rasch, Georg: *Probabilistic models for some intelligence and attainment tests*. 1. Auflage. Kopenhagen : Danish Institute for Educational Research, 1960

Rönz, Bernd: *Skript: Computergestützte Statistik I*. 2001. – Humboldt-Universität zu Berlin, Lehrstuhl für Statistik, Berlin

Rost, Jürgen: *Lehrbuch Testtheorie - Testkonstruktion*. 2. Auflage. Bern : Huber, 2004

Rousseeuw, Peter J. ; Leroy, Annick M.: *Robust Regression and Outlier Detection*. 1. Auflage. Wiley-Interscience, 2003

Strobl, Carolin: *Das Rasch-Modell*. 2. Auflage. Mering : Hampp, 2012

Suárez-Falcón, Juan C. ; Glas, Cees A. W.: Evaluation of global testing procedures for item fit to the Rasch model. In: *British Journal of Mathematical and Statistical Psychology* 56 (2003), Nr. 1, S. 127–143

Wainer, Howard ; Kiely, Gerard L.: Item clusters and computerized adaptive testing: A case for testlets. In: *Journal of Educational Measurement* 24 (1987), Nr. 3, S. 185–202

Wang, Wen-Chung ; Wilson, Mark: The Rasch Testlet Model. In: *Applied Psychological Measurement* 29 (2005), Nr. 2, S. 126–149

Wollenberg, Arnold L. van den: Two new test statistics for the rasch model. In: *Psychometrika* 47 (1982), Nr. 2, S. 123–140

Anhang

Deskriptive Statistiken

	n	m	ψ_{obj}	ψ_{dim}	ψ_{dep}
Minimum	100,0	10,00	0,0000033	0,0000463	0,0000071
1. Quantil	322,0	32,00	0,1251882	0,2455655	0,2461022
Median	550,0	54,00	0,2482903	0,4929474	0,4951508
Mittelwert	549,7	54,82	0,2487675	0,4984186	0,4969845
3. Quantil	777,0	78,00	0,3725568	0,7504744	0,7452634
Maximum	1000,0	100,00	0,4999279	0,9999941	0,9998618

Tabelle 7: Kennzahlen der Simulationsparameter über 9.998 Durchläufe

	$MAD_{\hat{\theta}}$	$AB_{\hat{\theta}}$	$MAD_{\hat{\delta}}$	$AB_{\hat{\delta}}$
Minimum	0,1947	0,0011	0,0431	0,000 E-00
1. Quantil	0,4306	0,0615	0,1223	0,081 E-16
Median	0,5230	0,1198	0,1617	0,177 E-16
Mittelwert	0,5173	0,1214	0,1657	0,215 E-16
3. Quantil	0,5989	0,1766	0,2059	0,303 E-16
Maximum	0,9850	0,3963	0,3726	1,649 E-16

Tabelle 8: Kennzahlen der Abweichungsmaße über 9.998 Durchläufe

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, 9. September 2013 Christoph Jaehrling